图书情报工作
第65卷 第3期 2021年2月

数字人文下的典籍深度学习实体自动识别模型构建及应用研究*

■ 杜悦[1] 王东波[1] 江川[1] 徐润华[2] 李斌[3] 许超[3] 徐晨飞[4]

[1] 南京农业大学信息科学技术学院 南京 210095 [2] 金陵科技学院人文学院 南京 210001
[3] 南京师范大学文学院 南京 210097 [4] 南通大学经济与管理学院 南通 226019

摘 要：[目的/意义]典籍是我国传统文化、思想和智慧的载体，结合数字人文的数据获取、标注和分析方法对典籍进行实体自动识别，对于后续应用研究具有重要意义。[方法/过程]基于经过自动分词与人工标注的25本先秦典籍构建古籍语料库，分别基于不同规模的语料库和Bi-LSTM、Bi-LSTM-Attention、Bi-LSTM-CRF、Bi-LSTM-CRF-Attention、Bi-RNN和Bi-RNN-CRF、BERT等7种深度学习模型，从中抽取构成历史事件的相应实体并进行效果对比。[结果/结论]在全部语料上训练得到的Bi-LSTM-Attention与Bi-RNN-CRF模型的准确率分别达到89.79%和89.33%，证实了深度学习应用于大规模文本数据集的可行性。

关键词：数字人文 深度学习 命名实体识别 先秦典籍
分类号：G255.1
DOI：10.13266/j.issn.0252-3116.2021.03.013

1 引言

自然语言处理在现代汉语的诸多领域都已取得较为丰硕的成果，但以先秦典籍为重要组成的古代汉语文本处理问题则亟待探索。在大数据时代的背景下，庞大的信息量使人们处理和理解信息的难度大增，传统的人文社会科学研究需要现代计算机技术的跨学科深层应用。国内对古文的利用和开发仍停留在传统的方法和模式上，古文典籍数据的大规模与利用深度不足的矛盾日益突出[1]。随着数字人文概念的出现，通过传统模式来对古籍进行开发利用的方法的不足之处愈发明显。古籍数字化不仅为古籍数据库、知识库的构建奠定了基础，而且为进行数字人文的探究提供了有力的数据支撑平台。通过对“数字人文”的研究，结合技术逻辑和人文逻辑，利用新的信息技术和跨学科方法构建可持续的、丰富的数据集和数据分析工具，可以实现对古籍的深度分析和挖掘[1]。作为中文信息处理的一个重要分支，古文词汇层级的处理基础任务包括自动分词、词性标注和命名实体识别等，其中命名实体识别的准确性和速度将影响后续研究的开展，其效果对于古汉语文本的深度挖掘有着重要意义。同时，对古文本中实体识别问题的探究不仅有助于数字人文技术应用领域，而且也有益于面向数字人文的古文语义知识库的构建。

目前进行命名实体识别的主流方法是统计与规则相结合的方法，这种方法在不同的语料上均取得了较高的准确性，其主要优势为通过抽象出文本的规则建立统计模型可以大幅度降低人工成本，但其依赖于专家经验和针对具体语料设计复杂的特征模板来提取特征，费时费力。深度学习模型利用已有的文本信息，对其上下文进行自动提取以掌握特征，进而探索其内部关系，有效缓解了传统方法存在的特征依赖与稀疏等

* 本文系国家自然科学基金面上项目“基于典籍引得的句法级汉英平行语料库构建及人文计算研究”（项目编号：71673143）和国家社会科学基金重大项目“基于《汉学引得丛刊》的典籍知识库构建及人文计算研究”（项目编号：15ZDB127）研究成果之一。
作者简介：杜悦（ORCID：0000-0001-7131-2325），本科生；王东波（ORCID：0000-0002-9894-9550），教授，博士生导师，通讯作者，E-mail：db.wang@njau.edu.cn；江川（ORCID：0000-0003-2436-9411），硕士研究生；徐润华（ORCID：0000-0003-0889-1808），讲师，博士；李斌（ORCID：0000-0002-7328-9947），副教授，博士；许超（ORCID：0000-0003-1051-5633），工程师，博士；徐晨飞（ORCID：0000-0002-9894-9550），讲师，博士研究生。
收稿日期：2018-11-25 修回日期：2020-10-30 本文起止页码：100-108 本文责任编辑：易飞

问题,目前在英文、现代汉语命名实体识别上均取得了令人满意的成果。古文语料较难获取且需要提前进行大量标注工作,且古文语料有其独特性质,在语法层面和句子长度上都与现代汉语和英文语料有一定差异,实体边界的划分很大程度上取决于分词的准确性,而先秦典籍作为我国传统文化、思想和智慧的最早载体更是如此,因此,针对古汉语文本的命名实体识别相对具有一定的挑战性。

实体识别是一个序列标注问题,主要是从非结构化的文本中提取人名、地名、时间等具有特定意义的事实信息。国外对命名实体的研究起步较早,与中文相比,英语单词之间存在明显的空格,字母具有大小写敏感性,因此国外命名实体的识别技术比国内成熟许多,准确率已达95%[2];C. Cherry 等[3]利用 Twitter 推文作为语料,使用词向量作为特征,大幅提高了 F1 值,达到了很好的识别效果;N. Peng 等[4]基于 LSTM 模型在自动分词上得到的较好结果,提出一种 LSTM 与 CRF 相结合的模型,F 值比之前单独使用 LSTM 模型的方法提高了5%;G. Lample 等[5]通过长短时记忆网络和基于转换的两种神经网络模型,从标注语料和未标注语料中提取特征,不借助任何特定语言知识或资源库,在英、德、西班牙、荷兰 4 种语言上均取得了目前最好的结果;此外基于卷积神经网络的多类分类方法也被应用于从电子病历中挖掘命名实体[6];混合深度神经网络(DNN)也被应用于命名实体识别,与条件随机场相比,在人名、地名和组织名的识别上均获得了显著提升[7]。国内目前使用深度学习模型进行实体识别的实践主要针对现代汉语文本,如人民日报语料(新闻)、微博语料(社会媒体)、化学药物名称(生物医学)等。相关研究如刘玉娇等[8]将深度学习方法应用于微博命名实体的识别,利用大量未标注的微博信息对自动编码器进行训练,获得抽象特征,随后将这些特征作为深度学习网络的输入,最后得出句子中每个字的类标;朱娜娜等[9]提出一种基于深度神经网络的表示学习方法,基于微博的数据特点,将候选图书名抽象为上下文连续的向量化表示,对微博内容中的图书名进行自动识别;陈佳浩[10]利用当前性能水平较好的卷积神经网络、循环神经网络等深度学习模型,针对在线文献中与人们日常生活关系最为紧密的食材名进行命名实体识别,取得了很好的效果。在中文地名识别方面,沈思等[11]利用循环神经网络方法,根据中文字和词的特点,重新定义了地名标注的输入和输出,基于深度学习方法提出了字级别的循环网络标注模型,准确率、召回率和 F 值均有明显提升;朱丹浩等[12]利用深度学习模型,完成了对中文机构名的识别。前述研究对于各种环境下的文本实体识别均取得了很好的效果,但其主要局限于现代文本,少数针对古汉语文本的实体识别研究也只针对某个方面,基于深度学习方法对古文中构成事件的实体进行抽取研究更是鲜少涉及。

本文利用 Bi-RNN、Bi-RNN-CRF、Bi-LSTM、Bi-LSTM-CRF、Bi-LSTM-Attention、Bi-LSTM-CRF-Attention、BERT 等 7 种深度学习模型,以《楚辞》《公羊传》《谷梁传》等 25 本先秦典籍为实验语料,对人名、地名、时间词 3 种可以构成历史事件的实体进行识别,并探究不同规模语料库对于先秦典籍命名实体识别效果的影响。

2 深度学习模型简介

2.1 循环类深度学习模型

循环神经网络(Recurrent Neural Network,RNN)是一种具有信息保存能力的神经网络结构,被广泛用于自然语言处理领域解决序列标注问题,可实现对长特征向量预测当前输出。在先秦典籍实体识别过程中,输入层是文本序列“大公封於管丘”,输出层是文本对应的标签“B-nr\E-nr\O\O\B-ns\E-ns”。与前馈神经网络相比,RNN 同一隐藏层之间的节点相互连接,使得隐藏层的输入由当前时刻的信息和之前时刻的信息共同组成,即判断“管”字向量的标签时,之前输入模型的文本序列“大公封於”均会对当前字向量的状态产生影响,共同决定其实体标签。其隐藏层和输出层中的值计算方法如下:

$$h(t)=f(W_1x(t)+W_2h(t-1)) \quad \text{公式(1)}$$

$$y(t)=g(W_3h(t)) \quad \text{公式(2)}$$

$$f(z)=\frac{1}{1+e^{-z}} \quad \text{公式(3)}$$

$$g(z_m)=\frac{e^{z_m}}{\sum_k e^{z_k}} \quad \text{公式(4)}$$

W_1 和 W_2、W_3 是在模型训练时被计算的连接权重,$f(z)$ 与 $g(z_m)$ 分别是 sigmoid 和 softmax 激活函数。

普通的循环神经网络只考虑一个方向的序列信息,而处理文本信息时另一个方向的序列信息同样重要,本文在这里将一个前向和一个后向的 RNN 上下叠加在一起,组成双向循环神经网络(Bi-RNN),可以从两个方向同时对句子进行学习,捕获整个句子的依赖关系。

循环神经网络(RNN)这种链式特征对于处理序列

化的数据具有很大的优势，理论上可以学习无限长的序列。但由于其记忆结构过于简单，Y. Bengio 等[13]发现模型对于比较长的输入存在梯度消失（vanishing gradient）的问题，梯度消失是影响 RNN 不能学习到无限长的序列的关键。作为 RNN 的一个变种，长短时记忆网络（LSTM，Long Short – Term Memory）引入细胞状态来存储信息而不是依靠单一的隐藏层[14]，通过三个门：输入门（input gate）控制当前时刻输入进入记忆单元的比例，忘记门（forget gate）决定当前记忆被忘记的比例，输出门（output gate）最终决定进入下一个神经网络单元的信息。

当“營”字向量进入模型，遗忘门决定先前文本序列“大公封於”将对当前词向量状态产生的影响，输入门保留部分向量信息，输出门则将“營”字向量信息与输入门、遗忘门输出信息整合后传入下一神经单元，对序列中的下一字向量“丘”的标签预测提供信息。长短时记忆网络的训练过程表示为数学公式如下：

$$i_t = \sigma(W_i * x_t + U_i * h_{t-1} + b_i) \quad 公式(5)$$

$$f_t = \sigma(W_f * x_t + U_f * h_{t-1} + b_f) \quad 公式(6)$$

$$o_t = \sigma(W_o * x_t + U_o * h_{t-1} + b_o) \quad 公式(7)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes tanh(W_c * x_t + U_c * h_{t-1} + b_c) \quad 公式(8)$$

$$h_t = O_t \otimes tanh(c_t) \quad 公式(9)$$

对于给定的含有 n 个词语的句子（$x_1, x_2, \cdots, x_n$），以《礼记》中“大公封于營丘”为例，首先将每个词转换为一个向量，长度为 d，然后通过模型来计算每个单词 t 的左上下文部分的表示向量，同样，为了获得右上下文部分的表示向量，也需要添加相应的信息，本文将前向 LSTM 与后向 LSTM 组合来获得文本左右部分上下文的表示向量 $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$，这一方法有效地包含了词语的所有上下文信息[15]，最终输入的文本序列经双向 LSTM 层预测可得到输出的标注序列。

在先秦典籍的实体识别任务中，本文使用“SBEIO”标签机制，如《礼记》中“大公封於營丘”对应标签为“B-nr”“E-nr”“O”“O”“B-ns”“E-ns”。从大规模标签的分布上看，标签不仅与自身的含义相关联，也受到上下文标签的影响。条件随机场模型已被广泛用于序列标注任务，取得了很好的效果，且在具体任务中，CRF 层聚焦于文本的句子级别而不是单一位置，也考虑了标签转移概率。为了进一步提高实体识别的准确性，我们将 CRF 与 Bi-LSTM 结合成为 Bi-LSTM-CRF 模型，在传统 LSTM 模型基础上加入整个句子的标签转移信息。该模型可以有效利用过去通过 LSTM 层输入的特征和通过 CRF 层输入的语句级别标记，并且新模型在利用上下文信息判定每一个词语的标签的同时，结合状态转移概率找到整个句子的最佳标签序列。

2.2 注意力机制类深度学习模型

注意力机制最早被应用到图像领域，随后被应用到自然语言处理领域，但目前并没有研究将注意力机制应用于先秦典籍的实体识别。在实体识别任务中，可通过注意力机制来获取篇章级信息，进而可以改善在一篇文章中相同词标签非一致性问题[15]，神经网络的每个节点通过注意力机制可以获得不同的概率权值，对目标词依赖度更高的节点对应更高的权重，以此优化模型性能。新的隐含状态由各个时刻的初始隐含状态通过加权和的形式计算得到，具体公式如下：

$$\vec{h} = \sum_{i=1}^{t} a_i h_i \quad 公式(10)$$

BERT 模型采用有较高建模能力的多层 Transformer 结构作为算法的主要框架，克服了 RNN 无法并行计算的缺点，结合注意力机制可以更全面地捕捉句子中的双向关系进而有效解决长依赖问题。在双向语言模型的基础上，BERT 加入了句子级别的连续性预测任务 NSP（next sentence prediction），在预训练时分两种情况生成训练文本，50% 的句子为语料中真正顺序相连的两个句子，剩下 50% 则从全部文本中随机选取一个片段拼接到第一个片段之后，预测输入 BERT 的两段文本是否为连续的文本。

2.3 传统机器学习模型

条件随机场是一种无向图模型，它计算给定输入节点条件下输出节点的条件概率，其公式如下：

$$P(y|x) = \frac{1}{z(x)} exp\left(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, x) + \sum_{i,k} \mu_k g_k(y_i, x)\right) \quad 公式(11)$$

其中 z(x) 是归一化因子：

$$z(x) = \sum_y exp\left(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, x) + \sum_{i,k} \mu_k g_k(y_i, x)\right) \quad 公式(12)$$

$$y' = argmax_y\ P(Y|X) \quad 公式(13)$$

CRF 的优势在于特征整合，在某些特征存在交叉的情况下依然能达到良好的性能[16]。

上述介绍的深度学习模型均以先秦典籍文本的向量表示形式为模型输入，采用双向网络架构以获取句子两个方向的文本信息。为进一步提升效果，本文引入 *LSTM* 中的门控机制与记忆单元来缓解 *RNN* 存在的梯度消失和梯度爆炸问题；将 *LSTM* 与 *CRF* 结合，判定每一个词的标签时通过状态转移矩阵结合整个句子的最佳标签序列；为体现相同的词向量在不同上下文句

子中的重要性,本文在神经网络架构中引入注意力机制,以期提升实体标签的分类结果。除此之外,将预训练模型 *BERT* 迁移至先秦典籍语料,模型输入不再是单一的文本 *Token* 嵌入,进一步探究将预训练阶段已获得的大量语言学知识应用于先秦典籍实体识别问题的可行性,也是本文研究的一个重点方向。

3 先秦典籍实体识别实验

3.1 语料库简介

本文使用的训练语料是经过手工分词和词性标注得到的《春秋左氏传》《诗经》《国语》等 25 本先秦典籍(具体典籍信息见表 1),是目前古文信息处理研究中涉及到的规模最大的语料,其涵盖了历史、典章制度、语言文字、政论、诗歌、军事等不同体裁和题材,比较完整地涵盖了先秦典籍的语言面貌,反映了先秦时期复杂的社会关系和文化现象,语料中需要被识别的人名、地名和时间词分别被标注为“nr”“ns”“t”。标注样例如下:

天王/n 使/v 劉定公/nr 勞/v 趙孟/nr 於/p 潁/ns ,/w 館/v 於/p 雒汭/ns。

表 1 语料具体信息

具体典籍	《楚辞》《孟子》《管子》《国语》《老子》《礼记》《墨子》《尚书》《诗经》《吴子》《孝经》《荀子》《仪礼》《周礼》《周易》《庄子》《左传》《商君书》《谷梁传》《韩非子》《吕氏春秋》《晏子春秋》《孙子兵法》

3.2 方法

首先在先秦典籍中使用“SBEIO”标签对文本进行单字序列标注,S、B、I、E、O 分别代表自身为先秦实体、先秦实体的左边界字、先秦实体的中间字、先秦实体的右边界字和非先秦实体。近几年,word2vec 的出现为利用深度学习模型处理自然语言问题提供了新的方向,并为从数字人文的角度深度挖掘古文本中所蕴含的潜在知识提供了极为有力的工具。word2vec 将词语表示成一个具有潜在语义信息的长度固定的低维向量,近似词语不仅在向量上具有相似性,它们之间还可以通过加减操作来获得词语之间的语义联系[17]。考虑到古籍文本的语言特性,使用字符向量作为深度学习模型的输入特征,以自动探索文本词语的潜在语义信息,而不需要基于规则和统计的方法根据外部特征对语料手工设置模板。其次,将分词标注后的语料按照 9:1 的比例以整句为单位随机拆分为训练语料与测试语料,并且为了探索语料库大小对实体识别效果的影响,将整个语料分为 1/4、1/2、3/4 与全部语料并进行 4 次对比实验。先秦典籍语料处理结果样例如表 2 所示:

表 2 先秦语料处理结果样例

序号	词语	词性/含义	标记
1	皐	nr/人名实体	S-nr
2	謂	v/动词	O
3	叔	nr/人名实体	B-nr
4	孫	nr/人名实体	E-nr
5	曰	v/无关词	O
6	:	w/无关词	O

3.3 实验设置

由于中央处理器无法满足神经网络模型在训练过程中所需的大量并行计算,因此,本文采用高性能 NVIDIA Tesla P40 图形处理器来训练神经网络,其处理能力比中央处理器快 60 倍以上,推理性能可达到 47 TOPS(万亿次/秒),保障其有足够的吞吐量和响应速度。本实验中使用的计算机配置如下:操作系统:CentOS 3.10.0;内存:256GB;显存:24GB;CPU:48 颗 Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz;GPU:6 块 NVIDIA® Tesla® P40。具体实验具体参数如表3 所示:

表 3 实验超参数设置

超参数	值
Bi-LSTM/Bi-RNN 层数	2
Hidden size	256
Learning rate	0.001
Batch-size	64
Dropout	0.5
Clip gradient	5

BERT 模型因其在语言模型与特征抽取架构上的不同,训练时需要的运算空间较大,与传统深度学习模型在参数设置上存在一些差异,具体实验参数如表 4 所示:

表 4 实验超参数设置

超参数	值
BERT 层数	2
Hidden size	128
Learning rate	2e-5
Batch-size	32
Train-epochs	3

4 先秦典籍实体识别实验结果及分析

4.1 实验结果

基于 Bi-RNN、Bi-RNN-CRF、Bi-LSTM、Bi-LSTM-Attention、Bi-LSTM-CRF、Bi-LSTM-CRF-Attention、BERT 模

型对“人名 nr、地名 ns、时间词 t”3 种命名实体进行自动识别，本文对比了不同语料库规模对于整个实验结果的影响。具体的评价指标为准确率 P(precision)、召回率 R(recall)，其计算公式如下：

$$P=\frac{\text{识别正确的实体}}{\text{识别正确的实体}+\text{被错误识别的实体}}*100\%$$

公式(14)

$$R=\frac{\text{识别正确的实体}}{\text{识别正确的实体}+\text{未被识别的实体}}*100\%$$

公式(15)

准确率与召回率相互影响，一般来说，前者较高则意味着后者较低，反之亦然，为了更客观地评价识别效果，不受单一指标影响，本文引入准确率与召回率的加权平均——调和平均值 F(F-Measure)，具体计算公式如下：

$$F=\frac{(\beta^2+1)*P*R}{(\beta^2*P)+R}=\frac{2*P*R}{P+R}\ (\text{当}\ \beta=1)$$

公式(16)

4.2　基于同规模语料库的不同模型效果分析

经过语料预处理及相应模型构建，基于全部语料库训练得到的 BERT 模型测试结果最佳，对人名、地名、时间词 3 种实体进行加权计算，得到如表 5 所示结果，绘制为柱状图后如图 1 所示。

表 5　全部语料各模型效果

模型	准确率(P)/%	召回率(R)/%	调和平均值(F)/%
Bi-LSTM	88.56	82.52	85.44
Bi-LSTM-Attention	89.79	82.66	86.08
Bi-LSTM-CRF	89.17	84.16	86.59
Bi-LSTM-CRF-Attention	88.31	83.28	85.72
Bi-RNN	85.88	81.37	83.57
Bi-RNN-CRF	89.33	81.33	85.14
BERT	86.21	86.67	86.44

从表 5 可以看出，在未进行任何人工提取特征的情况下，7 种不同的深度学习模型在先秦典籍实体识别上的应用均达到了较高的准确率，其中 Bi-LSTM-CRF 准确率达到了 89.17%，Bi-RNN-CRF 则达到了 89.33%，证明了深度学习模型应用于古汉语文本命名实体识别的可行性。观察表中其他结果，可以发现：

(1)与 Bi-RNN 模型相比，Bi-LSTM 模型结果的准确率、召回率及调和平均值均有提升，尤其是准确率从 85.88% 提升至 88.56%，较为显著。这一结果证明了相对于 RNN 存在的对于长序列文本输入的梯度消失问题，LSTM 引入记忆细胞与“门”机制保留历史信息

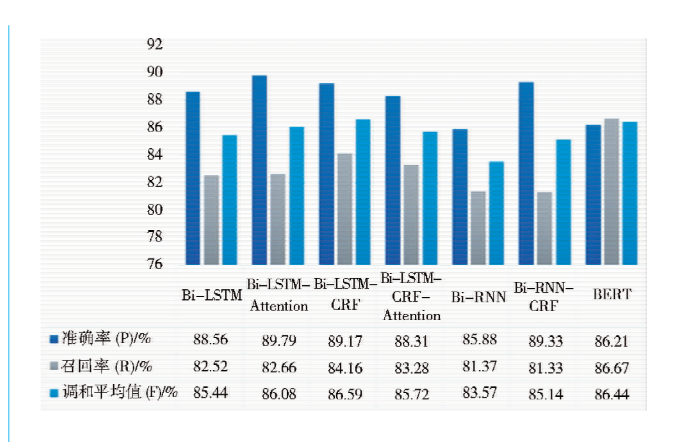


图 1　各模型在全部语料上的效果对比

的有效性，即在处理长距离依赖问题上的优越性。

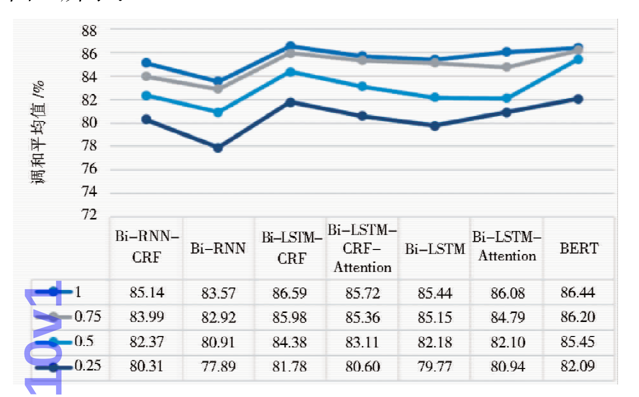
(2)在传统深度学习模型上加入 CRF 层后，Bi-RNN-CRF 相较于 Bi-RNN、Bi-LSTM-CRF 相较于 Bi-LSTM，其结果的准确率与调和平均值均有明显提升。证明了将 LSTM 与 CRF 结合，在针对先秦典籍实体识别的任务上具有一定的突出性。

(3)BERT 模型创新采用了基于注意力机制的多层双向 Transformer 架构以及在预训练中使用双向语言模型，其结果优于传统 RNN、LSTM 等模型结果，证实了该模型应用于大规模古文本命名实体识别任务的可行性。

(4)值得注意的是，在引入注意力机制后，传统模型识别效果并无明显提升，且 Bi-LSTM-CRF 实体识别的结果的相应指标发生了降低，与以往相关研究不符，也与我们的预估结果有很大不同。Bi-LSTM-CRF-Attention 相比 Bi-LSTM-CRF，其准确率、召回率与调和平均值均降低了接近 1%。经过文献调研，并对实验结果、语料特点进行分析后，本文尝试对这一结果做出解释：将注意力机制与 Bi-LSTM-CRF 模型结合应用于生物医学文本挖掘领域，效果获得了很好的提升[15]，但这一实验所用文本为生物医学领域英文文献，英文文献在语法与句法上与中文有很大的不同，且包含相同信息量时，英文文本语句更长、词汇更多，这也有利于注意力机制利用篇章级别的信息提升识别效果；在中文百科网站数据的实体识别任务中加入注意力机制也被证明是有效的[18]，但该实验数据利用爬虫爬取，是网络环境下的现代文语料，其篇章结构与注意力机制更为贴合。而我们所使用的先秦古籍语料，语句较现代文本更为精炼，句子中包含的平均词数较少，导致 Attention 层无法提取足够的特征信息，优势无法体现。

4.3 不同语料规模对实验结果的影响分析

为了探究在不同规模语料库下各模型的效果，本文将语料库规模分为 1/4、1/2、3/4 及全部，在 Bi-RNN、Bi-RNN-CRF、Bi-LSTM-CRF、Bi-LSTM-CRF-Attention、Bi-LSTM、Bi-LSTM-Attention、BERT 上进行实验，并将结果做对比，选取调和平均值 F 为评价指标，结果如图 2 所示：



	Bi-RNN-CRF	Bi-RNN	Bi-LSTM-CRF	Bi-LSTM-CRF-Attention	Bi-LSTM	Bi-LSTM-Attention	BERT
1	85.14	83.57	86.59	85.72	85.44	86.08	86.44
0.75	83.99	82.92	85.98	85.36	85.15	84.79	86.20
0.5	82.37	80.91	84.38	83.11	82.18	82.10	85.45
0.25	80.31	77.89	81.78	80.60	79.77	80.94	82.09

图 2 深度学习模型在不同规模语料库上的应用结果对比

基于以上展示结果，可以看到：

(1) 语料规模按每 1/4 语料库增长时，Bi-LSTM、Bi-RNN、Bi-LSTM-CRF 等模型实体识别结果均有明显提升。

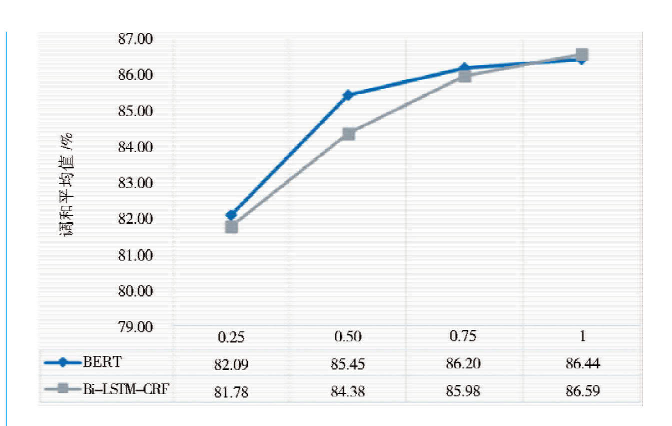
(2) Bi-RNN 与 Bi-LSTM-CRF 模型在不同规模语料库上应用，语料库规模从 1/4 增长至 1/2 时其结果的调和平均值有显著提升：提升指标均接近 3%，并且结果提升幅度随着语料库每次增加而递减，如从 3/4 规模增长至全部语料时效果只提升了 1% 左右。

为了直观展示 BERT 模型在不同规模语料上的应用效果，我们将 BERT 模型应用结果与传统深度学习模型中效果最好的 Bi-LSTM-CRF 作对比，以调和平均值为评价指标，结果见图 3。

基于以上展示结果，可以看到：

(1) 在全规模语料库上应用时，Bi-LSTM-CRF 模型效果略优于 BERT 模型，但在其他规模语料库上应用时 BERT 模型效果均为最佳，相较于传统深度学习模型以 LSTM 神经单元作为特征提取层，体现出 BERT 在采用多层双向 Transformer 作为特征提取器和双向语言模型上的创新性效果。

(2) BERT 模型在不同规模语料库上应用时，其调和平均值在语料库规模从 1/4 增长为 1/2、1/2 增长为 3/4 时有显著提高，可注意到语料库规模从 3/4 增长为全部时效果提升并不明显，BERT 模型应用于处理大规模文本时结果更加稳定，说明了 BERT 模型对于解决大规模数据集命名实体识别问题的优越性。



	0.25	0.50	0.75	1
BERT	82.09	85.45	86.20	86.44
Bi-LSTM-CRF	81.78	84.38	85.98	86.59

图 3 BERT 与 Bi-LSTM-CRF 在不同规模语料上的应用结果对比

基于上述结果，可得到以下分析：

语料库规模增加时，深度学习模型可以根据数据量的不断增加适当地扩展其规模，并从大量的文本数据中学习掌握更多上下文文本特征用于实体识别，从而有效防止在学习过程中过拟合与欠拟合情况的发生。并且本文使用 25 本先秦典籍中随机打乱并按比例分割的训练集与测试集进行训练，而先秦典籍中每一部其句法、词汇与写作风格都不完全一致，这也给我们后续使用深度学习模型进行实体识别带来了困扰，语料库规模的增加在使深度学习模型掌握更多文本信息的同时，也降低了整个训练集中噪声数据对整体效果的影响，可以更有效地对数据进行处理，证明了深度学习技术在处理大规模古籍数据集上的优越性，为先秦典籍的大规模、数字化深度开发提供了实践性方法，使得从数字人文的角度对古籍进行深度的文本挖掘和知识发现成为可能。

5 基于深度学习的典籍实体自动识别平台搭建

基于深度学习的典籍实体自动识别实验设计步骤较为复杂，如 25 部先秦典籍需要划分为不同规模语料库，再生成深度学习模型可识别的以整行形式存在的 tokens 并制作相应的特征模板，在对语料进行训练和测试后，还需要计算出其准确率 P、召回率 R 和调和平均值 F 3 个评价指标。为了便于实验展示，方便读者理解，本文基于 Bi-LSTM-CRF 构建了可视化最优深度学习典籍自动识别操作平台。

先秦典籍实体自动识别平台使用 Python 语言的第三方工具包 PyQt 进行开发。PyQt 是菲尔 · 汤普森开

发的 Python 语言的图形用户界面编程解决方案，可以在包括 UNIX、Windows 和 Mac 等的所有主要操作系统运行，成功继承了 Python 编程语言和 Qt 库，有 300 多个类和近 6 000 个函数和方法。相对于 wxPython、Tkinter 等图形库，PyQt 功能强大，可以使用“Qt Designer”或“Qt Creator”设计 UI 文件，从而简化了 UI 的设计布局等工作。

该平台主要包含两个部分，第一部分完成语料库构建，包括选择语料库规模和查看文本样例。第二部分是实体识别功能，包括生成特征模板与划分训练集和测试集、抽取实体。

构建语料库时，首先点击下拉框控件选择语料路径，之后选择构建语料库规模，点击“构建”按钮，平台自动完成 25 部先秦典籍对应规模语料库的构建，如图 4 所示。点击“查看样例”按钮即可查看语料库部分文本内容，如图 5 所示。



图 4　先秦典籍实体自动识别平台构建语料库功能截图



图 5　先秦典籍实体自动识别平台查看样例功能截图

点击“生成模板”按钮，平台对语料库自动按 9∶1 比例随机划分为以“train”和“test”命名的训练集和测试集文档（见图 6），随后自动调用深度学习典籍实体自动识别模型对 test 文档进行先秦典籍实体自动识别并分类显示，图 7 和图 8 分别为先秦典籍人名、地名实体。

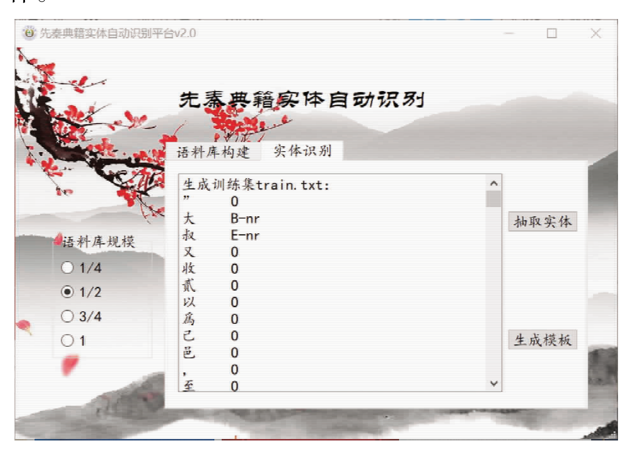


图 6　先秦典籍自动识别平台生成模板功能截图



图 7　先秦典籍实体自动识别平台抽取（人名）实体功能截图



图 8　先秦典籍实体自动识别平台抽取（地名）实体功能截图

6 结语

对古汉语文本进行人名、地名、时间词的实体自动识别对于古籍后续的相关数字人文研究起着重要作用。本文基于 Bi-RNN、Bi-RNN-CRF、Bi-LSTM、Bi-LSTM-CRF、Bi-LSTM-Attention、Bi-LSTM-CRF-Attention、BERT 等深度学习模型对 25 部先秦典籍进行了实体自动识别实验，并通过将语料规模按比例切分对比证明了基于深度学习模型对先秦古籍文本进行实体识别的可行性，以及深度学习模型在处理大规模文本数据集上的优越性。实验结果中 Bi-LSTM-Attention 与 Bi-LSTM-CRF 模型的准确率分别达到 89.79% 和 89.17%，证明了在传统深度学习模型上加入 CRF 与注意力机制的可行性以及 BERT 模型用于大规模文本集实体识别问题的优越性。同时本文的实验结果也表明，在英文和现代汉语的标准数据集上表现良好的某些深度学习模型不一定同样适用于先秦典籍的实体识别。

在后续研究中，将结合模型的整体表现性能，融合字词的统计特征来提高现有模型的评价指标，进一步探究将本实验涉及到的 25 本先秦典籍按照语料特点分组之后不同深度学习模型的实体识别效果，以期为不同古籍的实体识别工作提供一定参考。同时，随着数字人文技术、方法和理念的逐步推广和往纵深发展，如何把最新的人工智能中深度学习的技术与数字人文有机地融合起来也是未来的一个发展方向。

参考文献：

[1] 欧阳剑. 面向数字人文研究的大规模古籍文本可视化分析与挖掘[J]. 中国图书馆学报,2016,42(2):66－80.

[2] 谢韬. 基于古文学的命名实体识别的研究与实现[D]. 北京:北京邮电大学,2018.

[3] CHERRY C, GUO H. The unreasonable effectiveness of word representations for twitter named entity recognition[C]//Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Denver: Association for Computational Linguistics, 2015: 735－745.

[4] PENG N, DREDZE M. Improving named entity recognition for chinese social media with word segmentation representation learning[J]. arXiv preprint arXiv:1603.00786 2016:149－155

[5] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[J]. arXiv preprint arXiv:1603.01360 2016:260－270.

[6] DONG X, QIAN L, GUAN Y, et al. A multiclass classification method based on deep learning for named entity recognition in electronic medical records[C]// 2016 New York scientific data summit. New York: IEEE, 2016: 1－10.

[7] WANG G, CAI Y, GE F. Using hybrid neural network to address Chinese named entity recognition[C]//IEEE 3rd International conference on cloud computing and intelligence systems. Shenzhen: IEEE, 2015: 433－438.

[8] 刘玉娇,琚生根,李若晨,等. 基于深度学习的中文微博命名实体识别[J]. 四川大学学报(工程科学版),2016,48(S2):142－146.

[9] 朱娜娜,景东,薛涵. 基于深度神经网络的微博图书名识别研究[J]. 图书情报工作,2016,60(4):102－106,141.

[10] 陈佳浩. 基于深度学习的在线健康文献食材命名实体识别[D]. 广州:华南理工大学,2017.

[11] 沈思,朱丹浩. 基于深度学习的中文地名识别研究[J]. 北京理工大学学报,2017,37(11):1150－1155.

[12] 朱丹浩,杨蕾,王东波. 基于深度学习的中文机构名识别研究——一种汉字级别的循环神经网络方法[J]. 现代图书情报技术,2016(12):36－43.

[13] BENGIO Y, SIMARD P, FRASONI P. Learning long-term dependencies with gradient descent is difficult[J]. IEEE transactions on neural networks, 1994, 5(2):157－166.

[14] 周青宇. 基于深度学习的自然语言句法分析研究[D]. 哈尔滨:哈尔滨工业大学,2016.

[15] 杨培,杨志豪,罗凌,等. 基于注意机制的化学药物命名实体识别[J]. 计算机研究与发展,2018,55(7):1548－1556.

[16] 章成志,苏新宁. 基于条件随机场的自动标引模型研究[J]. 中国图书馆学报,2008,34(5):89－94.

[17] 张海楠,伍大勇,刘悦,等. 基于深度神经网络的中文命名实体识别[J]. 中文信息学报,2017,31(4):28－35.

[18] 唐敏. 基于深度学习的中文实体关系抽取方法研究[D]. 成都:西南交通大学,2018.

作者贡献说明：

杜悦：论文初稿撰写，模型训练；

王东波：论文设计与思路，论文修改，模型训练指导；

江川：模型训练协助；

徐润华：语料及古文知识提供；

李斌：语料及古文知识提供；

许超：语料及古文知识提供；

徐晨飞：相关研究文献整理。

Construction and Application of Entity Recognition Model Based on Deep Learning of Classics in Digital Humanities

Du Yue[1] Wang Dongbo[1] Jiang Chuan[1] Xu Runhua[2] Li Bin[3] Xu Chao[3] Xu Chenfei[4]
[1] College of Information and Technology, Nanjing Agricultural University, Nanjing 210095
[2] College of Humanities, Jinling University of Science and Technology, Nanjing 210001
[3] College of Literature, Nanjing Normal University, Nanjing 210097
[4] Economics and Management School of Nantong University, Nantong 226019

Abstract: [**Purpose/significance**] The classics are the carrier of Chinese traditional culture, thought and wisdom. Combining the methods of data acquisition, labeling and analysis of digital humanities, it is of great significance for the automatic entity recognition of classics for subsequent application research. [**Method/process**] The corpus was constructed based on 25 pre-Qin literature that have been automatically segmented and manually annotated, based on the corpus of different sizes and seven deep learning models of Bi-LSTM, Bi-LSTM-Attention, Bi-LSTM-CRF, Bi-LSTM-CRF-Attention, Bi-RNN, Bi-RNN-CRF and BERT, we extracted the corresponding entities that constituted historical events and compared their effects. [**Result/conclusion**] The accuracy of the Bi-LSTM-Attention and Bi-RNN-CRF models trained on all corpus reached 89.79% and 89.33%, respectively, confirming the feasibility of applying deep learning to large-scale text datasets.

Keywords: digital humanities deep learning named entity recognition Pre-Qin literature